Prompting4Debugging: Red-Teaming Text-to-Image Diffusion Models by Finding Problematic Prompts

Year: 2023 | Citations: 113 | Authors: Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, Wei-Chen Chiu

Abstract

Text-to-image diffusion models, e.g. Stable Diffusion (SD), lately have shown remarkable ability in high-quality content generation, and become one of the representatives for the recent wave of transformative AI. Nevertheless, such advance comes with an intensifying concern about the misuse of this generative technology, especially for producing copyrighted or NSFW (i.e. not safe for work) images. Although efforts have been made to filter inappropriate images/prompts or remove undesirable concepts/styles via model fine-tuning, the reliability of these safety mechanisms against diversified problematic prompts remains largely unexplored. In this work, we propose Prompting4Debugging (P4D) as a debugging and red-teaming tool that automatically finds problematic prompts for diffusion models to test the reliability of a deployed safety mechanism. We demonstrate the efficacy of our P4D tool in uncovering new vulnerabilities of SD models with safety mechanisms. Particularly, our result shows that around half of prompts in existing safe prompting benchmarks which were originally considered"safe"can actually be manipulated to bypass many deployed safety mechanisms, including concept removal, negative prompt, and safety guidance. Our findings suggest that, without comprehensive testing, the evaluations on limited safe prompting benchmarks can lead to a false sense of safety for text-to-image models.